Data Cleaning and EDA of Titanic Dataset

Madeline Albright

2024-11-14

knitr::opts\_chunk$set(echo = TRUE)  
## Load necessary libraries  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(tidyr)  
library(skimr)

## Warning: package 'skimr' was built under R version 4.3.3

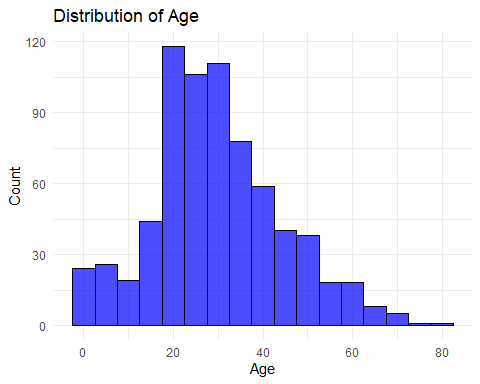
#install.packages("ggcorrplot")  
library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.3.3

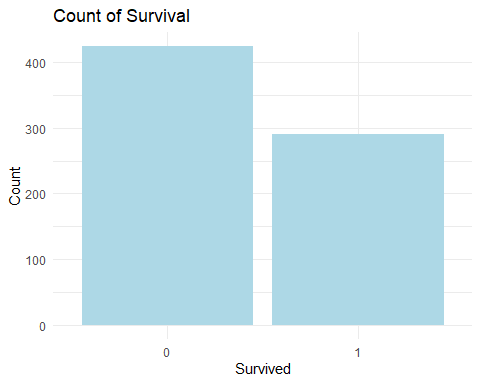
## Read the data  
data <- read.csv("C:\\Users\\USER\\Documents\\Futureinterns internship\\train.csv")  
  
## Clean the data  
data\_clean <- na.omit(data)  
data\_clean$Survived <- factor(data\_clean$Survived)  
data\_clean$Pclass <- factor(data\_clean$Pclass)  
data\_clean$Sex <- factor(data\_clean$Sex)  
data\_clean$Embarked <- factor(data\_clean$Embarked)  
summary(data\_clean)

## PassengerId Survived Pclass Name Sex   
## Min. : 1.0 0:424 1:186 Length:714 female:261   
## 1st Qu.:222.2 1:290 2:173 Class :character male :453   
## Median :445.0 3:355 Mode :character   
## Mean :448.6   
## 3rd Qu.:677.8   
## Max. :891.0   
## Age SibSp Parch Ticket   
## Min. : 0.42 Min. :0.0000 Min. :0.0000 Length:714   
## 1st Qu.:20.12 1st Qu.:0.0000 1st Qu.:0.0000 Class :character   
## Median :28.00 Median :0.0000 Median :0.0000 Mode :character   
## Mean :29.70 Mean :0.5126 Mean :0.4314   
## 3rd Qu.:38.00 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :80.00 Max. :5.0000 Max. :6.0000   
## Fare Cabin Embarked  
## Min. : 0.00 Length:714 : 2   
## 1st Qu.: 8.05 Class :character C:130   
## Median : 15.74 Mode :character Q: 28   
## Mean : 34.69 S:554   
## 3rd Qu.: 33.38   
## Max. :512.33

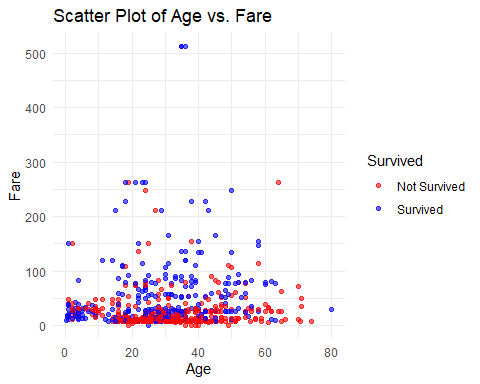
## Univariate Analysis  
# Histogram for Age  
ggplot(data\_clean, aes(x = Age)) +  
 geom\_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +  
 labs(title = "Distribution of Age", x = "Age", y = "Count") +  
 theme\_minimal()



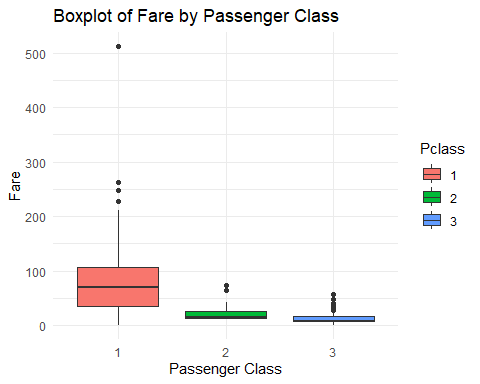
# Bar plot for Survived  
ggplot(data\_clean, aes(x = Survived)) +  
 geom\_bar(fill = "lightblue") +  
 labs(title = "Count of Survival", x = "Survived", y = "Count") +  
 theme\_minimal()



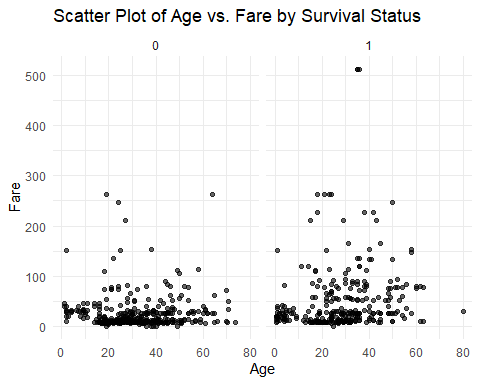
## Bivariate Analysis  
# Scatter Plot for Age vs. Fare  
ggplot(data\_clean, aes(x = Age, y = Fare)) +  
 geom\_point(aes(color = Survived), alpha = 0.6) +  
 labs(title = "Scatter Plot of Age vs. Fare", x = "Age", y = "Fare") +  
 theme\_minimal() +  
 scale\_color\_manual(values = c("red", "blue"), labels = c("Not Survived", "Survived"))



# Boxplot for Fare by Pclass  
ggplot(data\_clean, aes(x = Pclass, y = Fare)) +  
 geom\_boxplot(aes(fill = Pclass)) +  
 labs(title = "Boxplot of Fare by Passenger Class", x = "Passenger Class", y = "Fare") +  
 theme\_minimal()



## Multivariate Analysis  
# Facet Grid for Age vs. Fare by Survived  
ggplot(data\_clean, aes(x = Age, y = Fare)) +  
 geom\_point(alpha = 0.6) +  
 facet\_wrap(~ Survived) +  
 labs(title = "Scatter Plot of Age vs. Fare by Survival Status", x = "Age", y = "Fare") +  
 theme\_minimal()



# Correlation Matrix  
corr\_matrix <- cor(data\_clean %>% select(Age, Parch, Fare, SibSp), use = "complete.obs")  
ggcorrplot(corr\_matrix, lab = TRUE, title = "Correlation Matrix")

